Empirical Research in Vocational Education and Training
a SpringerOpen Journal

RESEARCH **Open Access**

Measurement of vocational competences: an analysis of the structure and reliability of current assessment practices in economic domains

Esther Winther and Viola Katharina Klotz*

* Correspondence:
viola.klotz@wiwi.upb.de
Chair of Business and Human Resource Education, University of Paderborn, Warbuger Straße 100, D-33098 Paderborn, Germany

Abstract

Background: Both fostering and measuring action competence remain central targets of vocational education and training research; adequate measurement approaches clearly are prerequisites for international, large-scale assessments. For the German Chamber of Commerce and Industry, competence assessments of industrial managers rely mainly on final examinations that attempt to measure not just knowledge but also action competence. To evaluate this test instrument, this article considers two questions: (1) Can the test assess action competence with validity, and (2) how reliable are the corresponding assessment results?

Methods: The study relied on statistical procedures (e.g., IRT scaling), applied empirically to a sample of 1,768 final examinations.

Results: As a result the current examination appears neither adequate nor accurate as an instrument to capture action competence.

Conclusions: We conclude that several improving steps have to be undertaken to improve the economic assessment.

Keywords: Vocational competences; Action competence; Item response theory (IRT) scaling; Competence structure; Test reliability

Background

Prospects and demand for adequate competence assessments

Explicit or implicit measures of vocational competence are relevant to many facets of vocational education and training (VET) and thus constitute an ever-growing research field. They pertain to national educational factors, such as relevant information and instruments for managing the quality of the vocational educational systems and developing adequate support programs, but increasingly, they also appear in international policy agendas. That is, international comparisons and acknowledgement of qualifications, as well as the encouragement of lifelong, informal learning, require adequate measurement concepts and innovative evaluation methods. To meet these multiple expectations, two major conditions must be fulfilled a priori (Klotz & Winther, 2012).

First, we require empirically confirmable competence models that encompass conceptual operationalizations of competences but also reveal a well-postulated theoretical structure that captures their empirical structure. From a scientific perspective, researchers seek empirical results related to the "true" structure of professional competences. From a political point of view, knowledge about the structure and comparability of competences is required to achieve large-scale assessments of VET, such as across Europe. In this context, compulsory education likely refers to a common curriculum of basic competences, such as literacy or numeracy, but the structure of competences within VET is more varied in content and therefore tends to be more complex. Thus VET content is heterogeneous not only between countries but also across different professions within nations (Baethge, Arends, & Winther, 2009) and even in specific workplaces (Billett, 2006). This abundant variation creates an ongoing dilemma for constructing generally valid competence tests. Uncertainty about the structure of competences also undermines international comparisons and the development of binding international agreements for consistent competence standards. Some scarce empirical research into the appropriate structure or model of competence suggests a content-based classification, such that item content exerts a characteristic influence on its difficulty. Other studies assume dimensionality based on different cognitive processing heuristics, which may determine response behaviors (Nickolaus, 2011; Nickolaus, Gschwendter, & Abele 2009; Nickolaus, Gschwendter, & Geißel 2008; Rosendahl & Straka, 2011; Seeber, 2008; Winther & Achtenhagen, 2009b, 2010).

Second, another necessary condition pertains to the reliability of the test results, that is, the certainty with which we can classify students according to a chosen test instrument. Neglecting this conditions poses serious risks, because people easily can be misclassified based on their test results, and such classification errors can have severe consequences for their future professional advancement.

With this study, we seek to evaluate both necessary conditions with respect to current testing efforts based on final examinations. Specifically, we describe how the German VET system currently operationalizes and measures competences in the economic domain. Empirical results obtained from a sample of 1,768 final examinations of industrial managers[a] reveal the extent to which German assessment instruments are qualified, in terms of their validity and reliability, to measure and classify students' economic action competence. This study, in accordance with a broader research program, seeks to develop and test a theoretical competence model and thereby improve current assessment practices. Its results thus offer guidelines for further development of the test instrument, as we discuss before concluding this article.

Conceptualization of final examinations

Action competence offers a constitutive element of the German vocational system and a significant topic of scientific and political discourse since the early 1980s, particularly in relation to the didactic implications of action regulation theory (Hacker, 1986; Kuhl, 1994a, 1994b; Volpert, 1983). In the mid-1990s, the Standing Conference of the Ministers of Education and Cultural Affairs (*Kultusministerkonferenz*) legally adopted the concept of action competence as a central target. Specifically and by law, students must be instructed in a way that enables them to *plan*, *execute*, and *monitor* an entire action process in a working environment. This concept appears largely heuristic but still must form the foundation for any test construction (BGBI, 2005 §5). In practice, these assessments come from the German Chamber of Commerce and Industry

(GCCI) and comprise both oral and written components. The oral part consists of a presentation and then a related expert discussion; it accounts for 30% of the assessment. The written examination comprises practical tasks pertaining to economics and social studies, as well as commercial management and control, together with situational tasks that take the form of case studies related to business processes. This last business processes section represents the most important assessment area, in terms of processing time (180 minutes) and weighting (40% of the final grade) (see Table 1). Therefore, this study focuses on this assessment component.

Recent commentary suggests that these test practices fail to give students sufficient room or potential to apply their knowledge to solve complex problems in a process-oriented working context (e.g., Haasler, 2007; Schmidt, 2000; Winther, 2010b) According to the GCCI (2009), the design of the business processes test component is intended to require test takers to model processes, undertake complex tasks, analyze business processes, and solve problems in an outcome- and customer-oriented way. To implement these goals, the test designers operationalized action competence as the three mutually exclusive process dimensions in Figure 1: planning, executing, and monitoring (GCCI 2009). Thus again, the business processes section seems particularly suitable for our empirical analysis of the structure of action competence.

If these process dimensions actually characterize a test situation, their solutions should require different sets of cognitive abilities of the test taker. In addition to this primary test conception, each item might be categorized according to four content domains: *marketing and distribution*, *acquisition*, *human resource management (HRM),* and *goods and services*. Such an alternative content-related model of competence measurement, as in Figure 2, appears in some other vocational assessments (Nickolaus, 2011; Nickolaus, Gschwendter, & Geissel 2008; Rosendahl & Straka, 2011; Seeber, 2008).

The content-related structure model for the economic domain reflects the previous curriculum of commercial schools, which were officially abolished in 1996, replaced by cross-disciplinary learning fields that sought to foster greater action competence.

Methods

Validity

Tests of validity determine if and to what extent a measurement actually measures the intended construct. This criterion comprises two facets. First, it describes the operationalization of a theoretical concept, together with its potential subdimensions and observable indicators, to determine if the focal approach offers a good measurement notion in relation to the latent trait. It therefore entails the translation of the latent trait into contents, and then the contents into reasonable measurement items, and in this sense, if refers to *content validity*. But even if an abstract concept is carefully operationalized, including all theoretical aspects and a reasonable item design, it

Table 1 Final Examination by the GCCI

Oral examination	Written examination
Presentation (10%)	Economics and social studies (10%)
Expert discussion (20%)	Commercial management & control (20%)
	Business processes (40%)

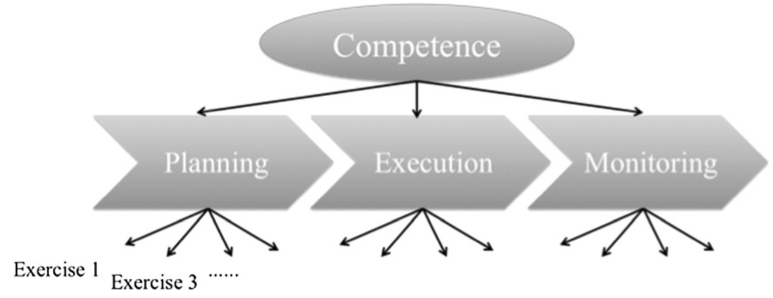

Figure 1 Procedural structure model.

remains possible that the theoretical concept simply does not exist in the real world—or at least not in the way assumed by the researcher. Second, to address the potential gap between theory and observed reality, validity assessments entail *construct validity* to determine if the postulated process and content structures arise from empirical test results.

Examination of content validity

Winther (2011) has analyzed the focal final examinations with regard to their objectivity and content validity. The results indicate systematic biases, due to nonuniform scoring during the correction process (see Table 2). With regard to content validity, Winther (2011) notes that a predominant part of the curriculum is dedicated to the *goods and services* domain (47% of the curriculum, about one-third of practical training), yet the proportion of content related to that topic in the test is rather small (21%). Thus, the test does not achieve representative validity. In particular, tasks related to modeling the processes of value creation and quantifiable production management are underrepresented, whereas the *marketing and distribution* content area appears overrepresented (38% of the final), in comparison with both its percentage of the curriculum (26%) and its practical relevance (25%).

Regarding *construct validity*, neither procedural nor content-based structures are clearly identifiable, perhaps due to the strong correction bias in the data (Winther, 2011). These results prompted a central re-correction of the examinations, such that the test results were compared, independent of the analyst, to gain unbiased data for further analyses of construct validity and reliability.

Examination of construct validity

Construct validity exists if the postulated process and content structures are actually reflected in empirical test results. To analyze theoretical structure models, most research relies on factor analytical approaches, though increasingly, multidimensional item response theory (IRT) models have grown in popularity (Hartig & Höhler, 2008). In accordance with this theory, a set of mathematical models describe, in probabilistic terms, the relationship between a person's response to an item and the level of a latent trait (e.g., Reeve & Fayers, 2005). Traditional approaches to measurement scales rely on averages or a simple summation of the test scores; IRT models instead reflect the assumption that the probability of solving an item depends on the test taker's latent trait or ability (i.e., θ_i = person parameter), combined with

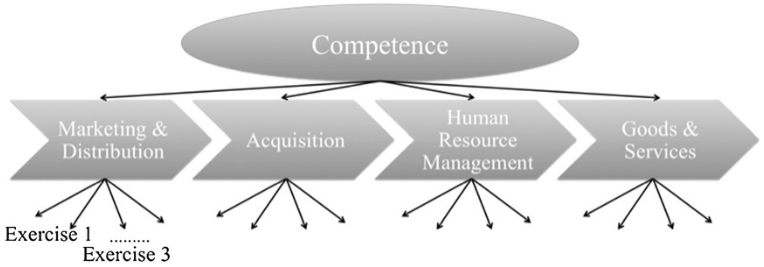

Figure 2 Content-related structure model.

the item difficulty (i.e., δ_i = item parameter). These two parameters relate negatively ($\theta_i - \delta_i$), because the probability of solving an item increases with the person's ability but decreases with greater item difficulty (Wright & Stone, 1979). This basic assumption can be formalized as a nonlinear function, namely, the item response function:

$$p(X_{vi} = x) = \frac{\exp(\theta_v - \delta_i)}{1 + \exp(\theta_v - \delta_i)} \qquad x = 0, 1. \tag{1}$$

It also can be depicted in an item characteristic curve, as in Figure 3.

For the analysis of the final examinations, we used IRT models because their traits and characteristics render them particularly suitable for this research goal[b]. However, a basic assumption underlying the application of parametric IRT models is that the model is appropriate for the data, which in turn demands the choice of the right model and an evaluation of model fit. The first consideration for choosing the right model is determining the number of item response categories. Only some structure models can model items with more than two response options, commonly referred to as polytomous items. In addition, the modeler must decide if another parameter, in addition to the item and person parameters (1PL model), can add to the level of item discrimination[c] (2PL model) or even if yet another parameter that reflects guessing effects[e] should appear in the model (3PL model) (Weiss & Davison, 1981). Although brevity considerations prevent us from describing all these models, we propose the specification scheme in Figure 4 to help render the decision process transparent and facilitate the search for an appropriate IRT model that can analyze the structure of competences in related research fields.

Competence measures often feature test instruments that contain polytomous, ordered item responses, such as the rating scale (Andrich, 1978), partial credit (Masters, 1982), and graded response (Samejima, 1969) models. Because competence, as measured by final examinations, seemingly constitutes a multidimensional concept, the

Table 2 Practical and curriculum relevance of examination contents

Items	Score (/100)	Practical learning (/25 months)	School hours (/600)	Curriculum weight
Marketing & Distribution	38	5-7 months	160 h	26.67%
Acquisition	20	5-7 months	80 h	13.33%
HRM	21	2-6 months	80 h	13.33%
Goods & Services	21	6-10 months	280 h	46.67%

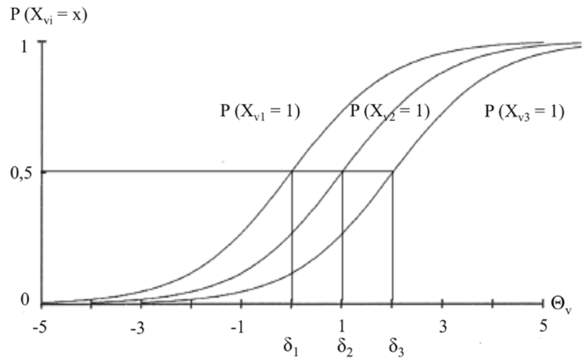

Figure 3 **Item characteristic curves for a dichotomous Rasch model with three items of varying difficulty (See Rost, 2004, p. 125; Winther, 2010a, p. 41).**

confirmation of its structure requires a multidimensional modeling approach. If competence tests contain items with various scales, as is likely in complex modeling situations, the partial credit model appears most appropriate. An advanced alternative also could take advantage of a 1PL model but still allow for varied scaling, that is, by fixing the discrimination parameter of the 2PL graded response model to equal 1 and thereby obtain the related 1PL model. The choice between these two models is somewhat arbitrary; both produce nearly identical results, albeit with slightly different parameterizations. Furthermore, this approach is easy to program using Mplus software, so this study adopts it to identify and evaluate whether the postulated theoretical structures appear in the final examination data. Accordingly, we allow for items with different numbers of response categories, as well as varying distances across response categories (e.g., Gibbons et al., 2007).

Examination of test reliability

The term “reliability” describes the replicability and thus the accuracy with which each item measures its intended trait. To assess a student’s expertise, a measure must have a strong probability of correctly classifying each student as possessing a certain competence value. For this analysis, we again applied an IRT standard. An important characteristic of IRT models is that they describe reliability, in terms of measurement precision, as a continuous function that is conditional on the values of the measured construct. It is therefore possible to model the test’s reliability for each individual value of competence for every test taker. The crucial appraisal criterion for a test’s reliability is measurement error, which arises because any measurement concept can include only a limited sample of the many possible items that constitute the measurement domain. The testing conditions also may vary, because factors other than student knowledge affect response behaviors, including both *student-specific* factors, such as mood, health, or individual differences in exposure to the tested content, and *situational* factors, such as distractions during the test, room temperature, and so forth (Kiplinger, 2008).

According to Fischer (1974), item precision can be depicted by item information curves (or functions), which indicate the range over the measurement construct in which the item discriminates best among individuals. The inverse of the squared

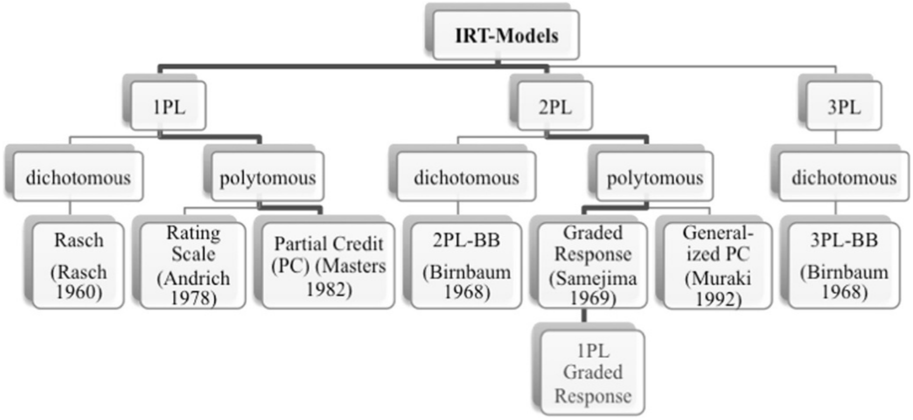

Figure 4 IRT specification scheme[f].

standard measurement error is equivalent to item information with respect to the latent trait (in our case, expertise). Thus,

$$I_i = \frac{1}{\sigma_i^2} \tag{2}$$

The higher the estimation variance, the less test information is available, and the lower the test's reliability (Ramsay, 1995):

$$Rel(\theta) = \frac{1}{1 + \frac{1}{I(\theta)}} \tag{4}$$

If the information is expansive, it is possible to identify a test taker whose true ability is at that level with reasonable precision.

Results and discussion

Results for the Test's validity

Testing both structures (i.e., process-oriented and content-related) within a single, integrated, 12-factor structure model was too unwieldy for the focal database, with only 35 items to distribute across dimensions. Numerically, the question of which theoretical model fits the real database best can be answered most effectively by so-called fit indices. In the test to confirm the processual structure model (M1 from Figure 1), we obtained poor values; this test concept does not appear valid for capturing competence. In contrast, the empirical evidence obtained for a school subject–oriented, content-

Table 3 Global fit indices for the procedural model (M1) and content-related structure model (M2)

Fit indices	*Cut-off criterion*	*M1*	*M2*
χ^2	≥ 0.05	0.000	0.000
Weighted Root Mean Square Residual	≤ 0.05	0.054	0.041
WRMR	≤ 0.9	2.036	1.663
Confirmatory fit index	≥ 0.95	0.782	0.957
Tucker-Lewis (1973) index	≥ 0.95	0.867	0.965

related measurement approach (M2) suggested good fit with the content structure for most items, as the comparison in Table 3 reveals.

To derive the content-related structure model, we used exploratory factor analysis. Specifically, to determine the number of factors, we combined a graphical scree test (Cattell, 1966) with a parallel analysis (Horn, 1965), using the MonteCarlo PA software, which offers a more objective approach for extracting factors. A five-factor solution emerged. We rotated the factor solution using oblique rotation method promax in SPSS, which is well suited to an analysis that allows for some correlation of factors (as can be assumed for the competence dimensions) and for very large data sets (as is the case for the final examinations). During this analysis, the data freely generated the postulated contend-related structure model, together with the predicted parameters of the model. In the only empirical difference, the contents of the academic subjects *marketing and distribution* split empirically into two domains (*marketing* and *distribution*), as we show in Figure 5.

Thus, the content-related structure model supports the validity of 21 of the 35 items with regard to their effectiveness for measuring differences in the abilities of test takers. However, the concept measured is not actually action competence, as intended, but rather content-related, technical knowledge, in an expertise-related sense. If we also consider the content of items not represented in this structure, we note that these abilities are characterized by their relatively transferable, contextualized nature and often involve calculations.

Results for the Test's reliability

Using IRT-standard, the amount of information can be computed for each ability level on a test's ability scale (Baker, 2001). We show the results for the final examinations data in Figure 6.

The information function for the test reaches its maximum for persons with an approximately average competence level. That is, near this area, it is possible to estimate, very precisely, test takers' true level of expertise (reliability = .88). Farther from this maximum though, the test's estimation precision decreases rapidly. Students with relatively high ability, who are located in the positive space, reveal a lower but still sufficient information value. In contrast, students with below-average expertise get estimated with an information value tending to 0. Because the test information reflects the sum of individual item information at a given ability level,

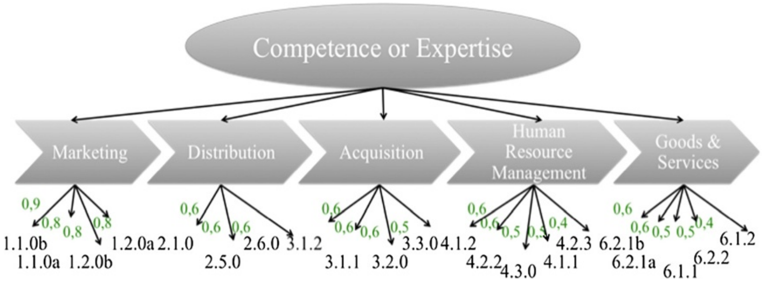

Figure 5 Empirically generated content-related structure model.

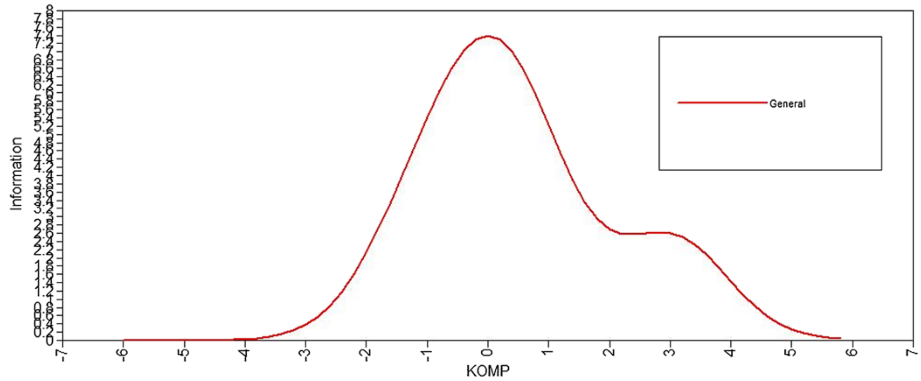

Figure 6 Test reliability for the GCCI final examinations.

the amount of information also is defined at the item level. The test provides many measurement items related to an average ability level, along with some items to measure high ability levels, but it features few easy items designed to measure low levels of expertise. Therefore, the GCCI final examinations cannot effectively differentiate test takers with low versus very low ability.

However, this gap does not necessarily cause problems. Some tests are constructed explicitly to differentiate students precisely at a specific, crucial point. That is, we need to consider the specific purpose of any particular test instrument to assess its reliability. The primary purpose of the final examinations is to regulate access to the industrial management profession, such that test takers are separated simply into those who pass the test, and thus receive certification to enter the professional community, and those who do not. Annually, approximately 95% of test takers pass,[f] so the most important separation point must fall far below an average competence level. Yet the amount of test information available in this range tends toward zero, so students have been quasi–blindly classified into the crucial "passed" or "failed" categories. This lack of reliability in final examinations not only infringes on statistical test standards but also has severe implications for the professional development and life of a vast number of students.

Conclusions

The evaluation of the validity of action competence provided by this article reveals that the assessment entails not the intended, process-oriented structure but rather a fractured, subject-specific, content structure. This content-related structure model reflects a previous, officially abolished teaching structure and curriculum, which makes it quite surprising that this conceptualization still dominates the test. The instrument may be partially valid for assessing subject-specific content—that is, the expertise of a student in several subjects—but it cannot capture true action competence.

Furthermore the items do not demonstrate reliability in their ability to depict the expertise of a student in several subjects. The empirical results pertaining to the structure of vocational competence are coherent with studies in other vocational areas that similarly suggest the high relevance of subject-related domains in the structuring of professional competence measures and their frequent influence on item difficulty (e.g., Nickolaus, Gschwendter, & Abele, 2009; Seeber, 2008). However, for measuring competence acquired in VET, this approach seems insufficient. If action competence is not to devolve into simply a buzzword, the concept must be salient and manifest in final examinations. In particular, newly developed and implemented assessment practices must capture students' skills in thinking and reasoning effectively and solving complex problems autonomously, on the basis of constructivist theory (Gijbels et al., 2006; Pellegrino et al., 2001).

Finally, with regard to the accuracy with which the final examination distinguishes and classifies students, we find that it does not provide enough items to measure under average competence levels accurately. The poor reliability limits true classifications of learning outcomes, because students who have been classified as failures, and who are therefore denied certain positions within the professional community, easily could be misclassified. The informative value and explanatory power for the GCCI test instrument thus are low.

Because the current examination appears neither adequate nor accurate as an instrument to capture action competence, we propose improving the foundational conceptualization of the test by

1. Designing more items pertaining to the "acquisition" and "goods and services" content areas.
2. Offering adequately authentic and complex test situations, such that the process-oriented, situated item setting aims to model real-life, authentic situations (Shavelson, 2008).
3. Forming a vertical competence structure based on cognitive dimensions and developing situations with varying complexity, to test different action competence qualities and increase the interpretability of the IRT test scores (i.e., criterion-based assessment).
4. Designing more easy items, to achieve greater reliability at the most crucial separation point of the test.
5. Adopting a competence model that better depicts the development of competence throughout the learning process, moving from general competences (domain-related) to more specific competence components (domain-specific) (Winther, 2010a; Winther & Achtenhagen, 2008), focused on work requirements in specific occupations to stimulate company operations across departments and their specific economic features (Winther 1 & Achtenhagen, 2009a).

By incorporating such aspects into the final examination, the GCCI could make its assessment instrument more valid and move it beyond the current focus on component skills and discrete bits of knowledge, to encompass the more complex aspects of student achievement (Pellegrino et al., 2001). Furthermore, such a test structure might offer more information about the level of competence students actually acquire and concrete starting points for developing support measures to improve their learning process. Results from initial tests of this novel examination approach will be ready in late 2013.

Endnotes

[a] The data were acquired from six headquarters of the German Chamber of Commerce and Industry: Luneburg, Hanover, Frankfurt on the Main, Munich, Saarland, and Nuremberg.

[b] Steyer and Eid (2001) note that a missing correlation between different error terms—as assumed in classical test theory—implies unidimensionality. In a probabilistic approach, this assumption disappears though, so (multi)dimensionality is explicitly confirmable. In empirical terms, the identified competence dimensions are sufficiently independent in their correlative cohesion (e.g., Hartig & Klieme, 2006).

[c] Item discrimination refers to an item's ability to differentiate among people at different levels along a particular trait continuum (Birnbaum 1968).

[d] Guessing effects describe a respondent's probability of getting a question correct, simply by chance.

[e] For the exact characteristics of each model, see Embretson and Reise (2000).

[f] Acquired from statistics for Munich and Upper Bavaria.

Competing interests

The authors hereby declare that they have no competing interests.

Author's contributions

Both authors contributed equally to this work. EW designed the study; VK analysed the data; both authors wrote and discussed the manuscript at all stages. Both authors read and approved the final manuscript.

Acknowledgement

This article arose from the subproject "Competence-oriented assessments in VET and professional development" (Wi 3597/1-1 and Wi 3597/1-2), within the framework of the priority programme "Competence Models for Assessing Individual Learning Outcomes and Evaluating Educational Processes" (SPP 1293) of the German Research Foundation (DFG).

Received: 12 June 2013 Accepted: 14 June 2013
Published: 16 July 2013

References

Andrich D (1978) A rating formulation for ordered response categories. Psychometrika 43:561–574

Baethge M, Arends L, Winther E (2009) International large-scale assessment on vocational and occupational education and training. In: Oser F, Renold U, John EG, Winther E, Weber S (eds) VET boost: Towards a theory of professional competences. Essays in honor of Frank Achtenhagen. Sense Publishers, Rotterdam

Baker F (2001) The basics of item response theory. University of Maryland, College Park, MD

Berufsbildungsgesetz (BGBl) (2005) vom 23. März 2005, in Kraft getreten am 1

Billett S (2006) Work, change and workers. Springer, Dordrecht

Birnbaum A (1968) Some latent trait models and their use in inferring an examinee's ability. In: Lord FM, Novick MR (eds) Statistical theories of mental test scores. Addison-Wesley, Boston, MA

Cattell RB (1966) The scree test for the number of factors. Multivariate Behavioral Research 1:245–276

Embretson SE, Reise SP (2000) Item response theory for psychologists. Lawrence Erlbaum Associates Publishers, Mahwah, NJ

Fischer GH (1974) Einführung in die Theorie psychologischer Tests. Huber, Bern

German Chamber of Commerce and Industry (GCCI) Aufgabenstelle für käufmännische Abschluss und Zwischenprüfungen (AKA) (Hrsg.) (2009) *Prüfungskatalog für die IHK-Abschlussprüfungen*. Vol. 3. , Auflage: Nürnberg

Gibbons R, Bock D, Hedeker D, Weiss DJ, Segawa E, Bhaumik DK, Kupfer DJ, Frank E, Grochocinski VJ, Stover A (2007) Full-information item bifactor analysis of graded response data. Applied Psychological Measurement 31:4

Gijbels D, Van De Watering G, Dochy F, Van Den Bossche P (2006) New learning environments and constructivism: The students' perspective. Instructional Science 34:213–226

Haasler B (2007) Anregungen zur Prüfungspraxis in der deutschen dualen Berufsausbildung aus der Perspektive der gewerblich-technischen Berufsausbildungsforschung. In: Grollmann P, Luomi-Messerer K, Stenström M-L, R Tutschner (Hrsg.) (eds) Praxisbegleitende Prüfungen und Beurteilungen in der Beruflichen Bildung in Europa. Bd. 18 Bildung und Arbeitswelt, Wien, Berlin, pp 193–220

Hacker W (1986) Arbeitspsychologie. Psychische Regulation von Arbeitstätigkeiten. Huber, Bern

Hartig J, Höhler J (2008) Representation of competences I multidimensional IRT. Models with within-item and between-item multidimensionality. Zeitschrift für Psychologie 216(2):89–101

Hartig J, Klieme E (2006) Kompetenz und Kompetenzdiagnostik. In: Schweizer K (ed) Leistung und Leistungsdiagnostik. Springer, Berlin

Horn JL (1965) A rationale and test for the number of factors in factor analysis. Psychometrika 30:179–185

Kiplinger L (2008) Reliability of large scale assessment and accountability systems. In: Ryan KE, Shepard LA (eds) The future of test-based educational accountability. Routledge, New York

Klotz VK, Winther E (2012) Kompetenzmessung in der kaufmännischen Berufsausbildung: Zwischen Prozessorientierung und Fachbezug. Eine Analyse der aktuellen Prüfungspraxis. In: *bwp@ Berufs - und Wirtschaftspädagogik – online, 22*

Kuhl J (1994a) A theory of action and state orientation. In J. Kuhl, & J. Beckmann (Eds.) *Volition and personality: Action vs. state orientation*. Hogrefe & Huber, Seattle

Kuhl J (1994b) Action and state orientation: Psychometric properties of the action control scales. In J. Kuhl, & J. Beckmann (Eds.) *Volition and personality: Action vs. state orientation*. Hogrefe, Göttingen

Masters GN (1982) A Rasch model for partial credit scoring. Psychometrica 47:149–174

Nickolaus R (2011) Die Erfassung fachlicher Kompetenz und ihrer Entwicklungen in der beruflichen Bildung - Forschungsstand und Perspektiven. Stationen empirischer Bildungsforschung: Traditionslinien und Perspektiven:331–351

Nickolaus A, Gschwendter T, Geißel B (2008) Modellierung und Entwicklung beruflicher Fachkompetenz in der gewerblich-technischen Erstausbildung. Zeitschrift für Berufs-und Wirtschaftspädagogik 104:48–73

Nickolaus R, Gschwendter T, Abele S (2009) Die Validität von Simulationsaufgaben am Beispiel der Diagnosekompetenz von Kfz-Mechatronikern. Vorstudie zur Validität von Simulationsaufgaben im Rahmen eines VET-LSA. Abschlussbericht für das Bundesministerium für Bildung und Forschung, Stuttgart

Pellegrino JW, Chudowsky N, Glaser R (2001) Knowing what students know: The science and design of educational assessment. National Academy Press, Washington, DC

Ramsay JO (1995) *TestGraf. A program for the graphical analysis of multiple choice test and questionnaire data* [Manual and Software]. Author, Montreal

Reeve BB, Fayers P (2005) Applying item response theory modelling for evaluating questionnaire item and scale properties. In: Fayers P, Hays RD (eds) Assessing quality of life in clinical trials: methods of practice, 2nd edn. Oxford University Press, New York

Rosendahl J, Straka GA (2011) Kompetenzmodellierungen zur wirtschaftlichen Fachkompetenz angehender Bankkaufleute. Zeitschrift für Berufs- und Wirtschaftspädagogik 107(2):190–217

Rost J (2004) Lehrbuch Testtheorie und Testkonstruktion. Huber, Bern

Samejima F (1969) Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph Supplement 34(4):100–114

Schmidt JU (2000) Prüfungen auf dem Prüfstand – Betriebe beurteilen die Aussagekraft von Prüfungen. Berufsbildung in Wissenschaft und Praxis 29(5):27–31

Seeber S (2008) Ansätze zur Modellierung beruflicher Fachkompetenz in kaufmännischen Ausbildungsberufen. Zeitschrift für Berufs- und Wirtschaftspädagogik 104:74–97

Shavelson RJ (2008) Reflections on quantitative reasoning: an assessment perspective. In B. L., Madison, & L. A., Steen. (Eds.) *Calculation vs. context: Quantitative literacy and its implications for teacher* education. Mathematical Association of America, Washington, DC

Steyer R, Eid M (2001) Messen und Testen. Springer, Berlin

Tucker LR, Lewis C (1973) The reliability coefficient for maximum likelihood factor analysis. Psychometrica 38:1–10

Viola Katharina K, Esther W (2012) Kompetenzmessung in der kaufmännischen Berufsausbildung: Zwischen Prozessorientierung und Fachbezug. Eine Analyse der aktuellen Prüfungspraxis *bwp@ Berufs- und Wirtschaftspädagogik - online, 22*

Volpert W (1983) Handlungsstrukturanalyse als Beitrag zur Qualifikationsforschung. Pahl-Rugenstein, Köln

Weiss DJ, Davison ML (1981) Test theory and methods. Annu Rev Psychol 32(1):629–658

Winther E (2010a) Kompetenzmessung in der beruflichen Bildung. Bertelsmann, Bielefeld

Winther E (2010b) Kompetenzen messen – Zur Notwendigkeit methodologischer und quantitativer Standards im Rahmen beruflicher Kompetenz. Zeitschrift für Berufs und Wirtschaftspädagogik 106(3):128–137

Winther E (2011) Kompetenzorientierte Assessments in der beruflichen Bildung – Am Beispiel der Ausbildung von Industriekaufleuten. Zeitschrift für Berufs- und Wirtschaftspädagogik 107(1):33–54

Winther E, Achtenhagen F (2008) Kompetenzstrukturmodell für die kaufmännische Bildung. Adaptierbare Forschungslinien und theoretische Ausgestaltung. Zeitschrift für Berufs- und Wirtschaftspädagogik 104(4):511–538

Winther E, Achtenhagen F (2009a) Measurement of vocational competences—a contribution to an international large-scale assessment on vocational education and training. Empirical Research in Vocational Education and Training 1:88–106

Winther E, Achtenhagen F (2009b) Skalen und Stufen kaufmännischer Kompetenz. Zeitschrift für Berufs- und Wirtschaftspädagogik 105(4):521–556

Winther E, Achtenhagen F (2010) Berufsfachliche Kompetenz: Messinstrumente und empirische Befunde zur Mehrdimensionalität beruflicher Handlungskompetenz. Berufsbildung in Wissenschaft und Praxis 1:18–21

Wright BD, Stone MH (1979) Best test design. MESA Press, Chicago

doi:10.1186/1877-6345-5-2
Cite this article as: Winther and Klotz: **Measurement of vocational competences: an analysis of the structure and reliability of current assessment practices in economic domains.** *Empirical Research in Vocational Education & Training* 2013 **5**:2.